RAG

BOBER

Сервис поиска параграфов

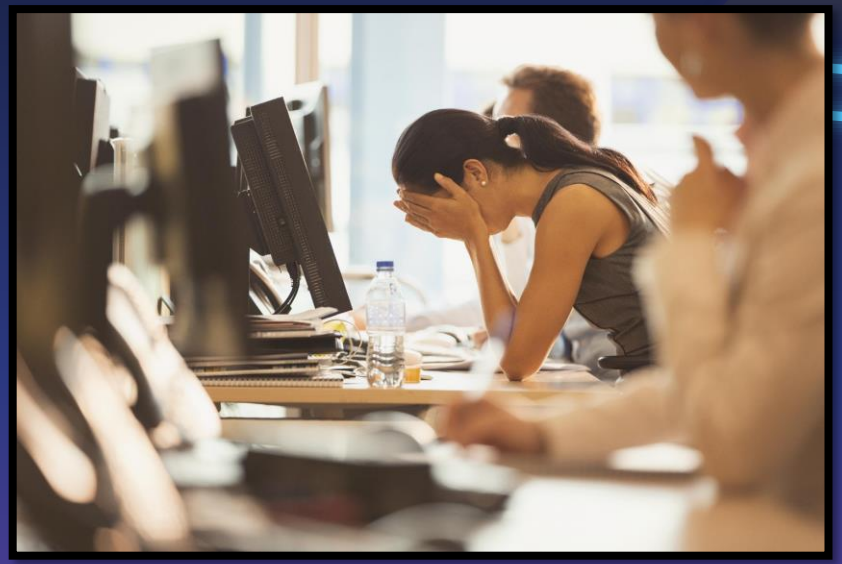
Кадочникова Маргарита	Университет ИТМО	Прикладная математика и информатика
Лаврентьев Степан	РГУ нефти и газа (НИУ) имени И.М. Губкина	Автоматика и вычислительная техника
Чикалева Юлия	Университет ИТМО	Веб-технологии
Хабаров Никита	Университет ИТМО	Информационные системы и технологии
Саркисян Арсен	Университет ИТМО	Химия и искусственный интеллект

КЕЙС 3
Сервис поиска параграфов

Сейчас **много времени сотрудников компании уходит на обработку запросов**, связанных с внутренними процедурами.

Для этого требуется:
- зафиксировать контекст запроса
- понять суть вопроса
- определить, где искать ответ на вопрос.

Запросы требуют сложного анализа и непростых ответов.



Ключевые метрики

Средняя зарплата сотрудника: **60 000 руб./мес.**

Время на ручной поиск: 3-4 часа ежедневно

Стоимость непродуктивного времени: 375 руб./час (60 000 / 160 рабочих часов в месяц)

46 875 руб.

В среднем за месяц с одного сотрудника

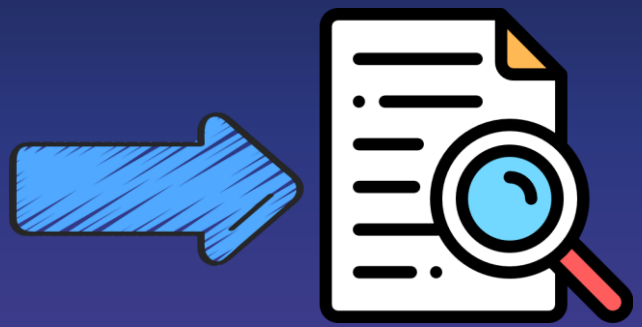
562 500 руб.

В среднем за год с команды из 12 человек

Устаревшие методы поиска информации снижают производительность компании

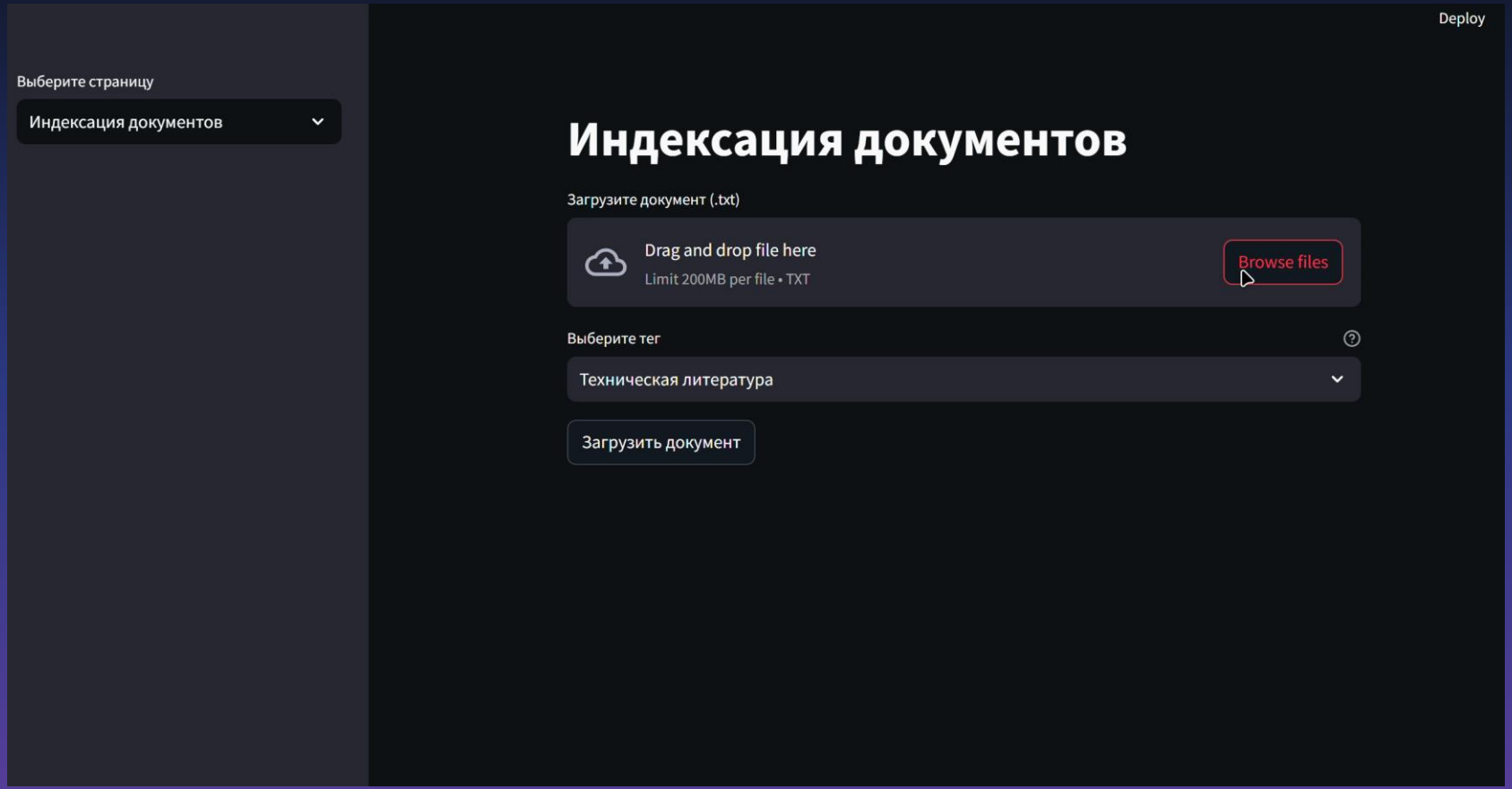
РЕШЕНИЕ

Загрузка

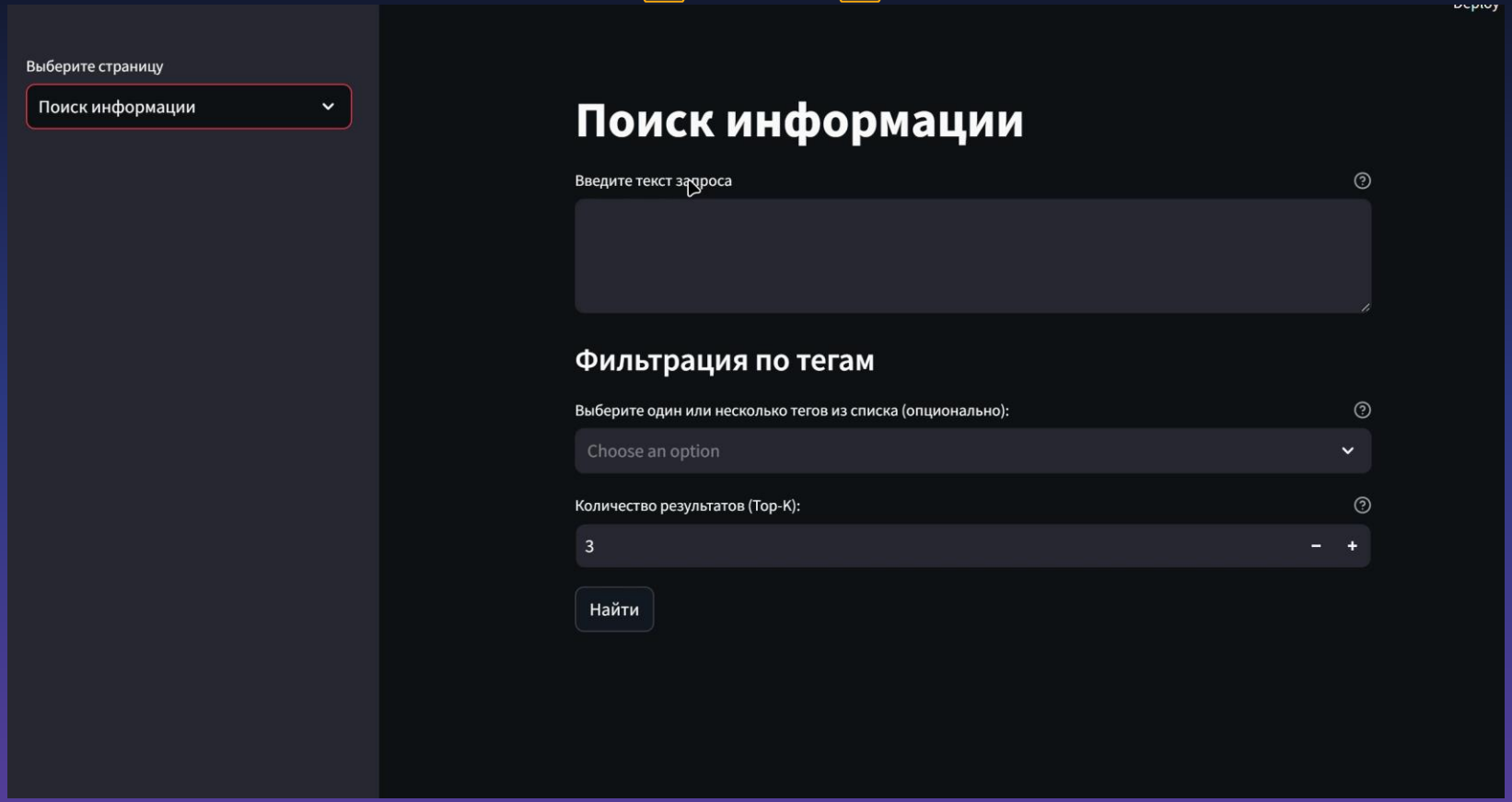


Поиск

Результат



Страница индексации документов: загрузка файлов, присвоение тегов



Страница поиска: ввод запроса, фильтрация по тегам
Отображение результатов

РЕШЕНИЕ

01

Поддержка русского и английского языков

02

Семантический поиск с использованием векторных эмбеддингов

03

Быстрый доступ к релевантной информации

Архитектура решения

FRONTEND

- Многостраничное приложение
- Индексация документов и поиск информации
- Загрузка документов, теги, поисковые запросы

BACKEND

- FastAPI
- Интеграция с HuggingFace Transformers (sentence-transformers/all-MiniLM-L6-v2)
- Связь с Weaviate

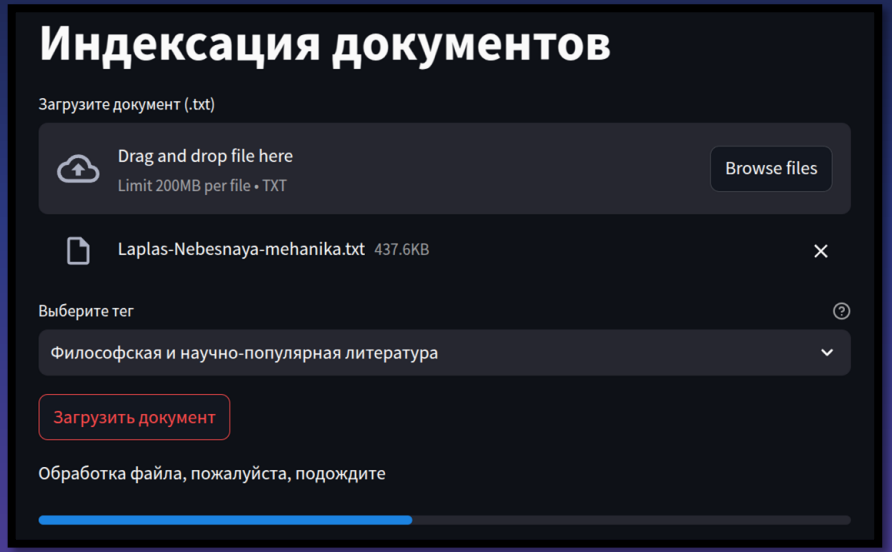
BD

- Векторная БД для хранения и поиска Weaviate.
- Гибридный поиск (текст + векторы)
- Фильтрация по метаданным

Архитектура решения. Фронтенд.

FRONTEND

- Streamlit, легкий в использовании интерфейс с возможностью загрузки документов, выбора тегов и фильтрации поиска.
- Индексация документов и поиск информации
- Загрузка документов, теги, поисковые запросы

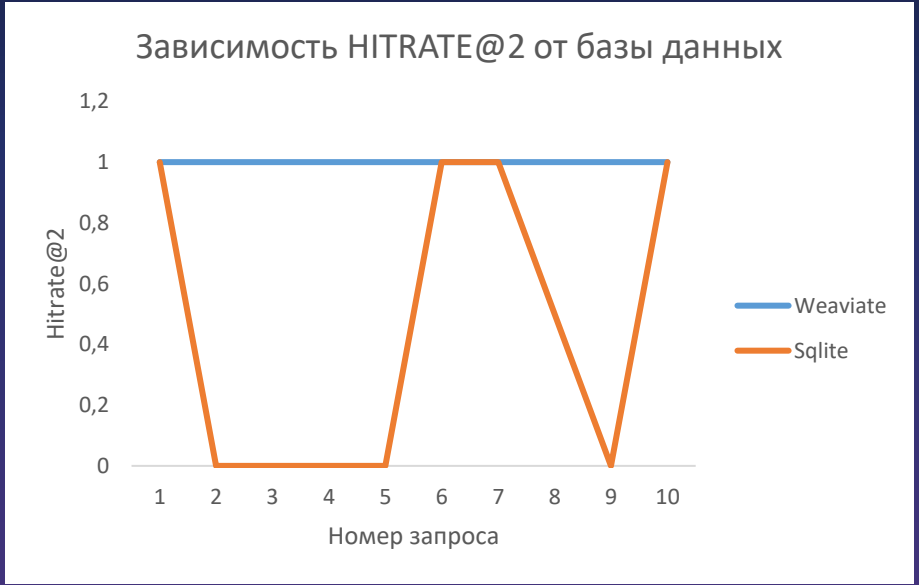


Архитектура решения. База данных.

Оптимизирован для работы с векторными данными, используется модель sentence-transformers/all-MiniLM-L6-v2

Встроенные методы гибридного поиска (текст + вектор)

Лучшая масштабируемость при больших объёмах данных



Архитектура решения. Бэкенд.

Эндпоинт /indexing

Добавление документов в базу:

1. Проверить подключение к базе.
2. Сохранить текст, метаданные и эмбеддинги в коллекции Data_base_paragraphs.
3. Вернуть: *{"message": "Documents indexed with embeddings"}*.

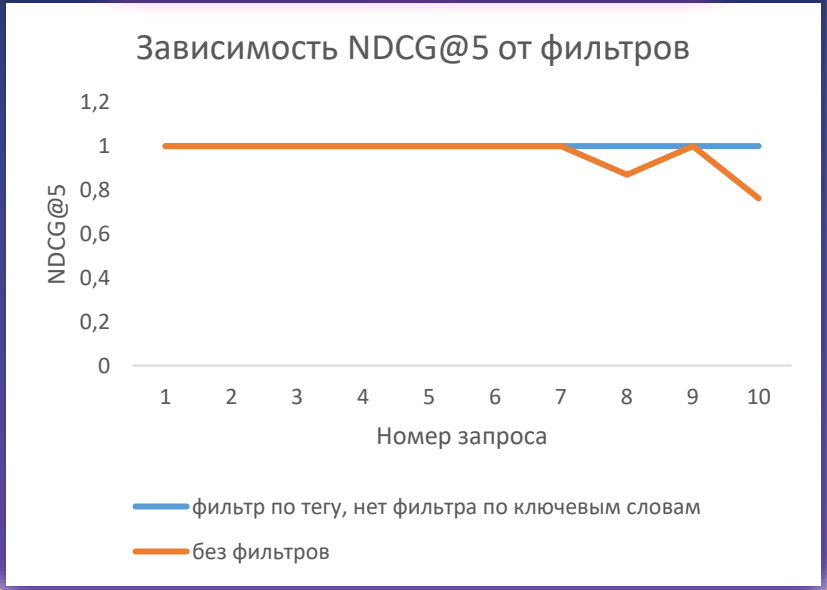
Эндпоинт /searching

Гибридный поиск (текст + эмбеддинги):

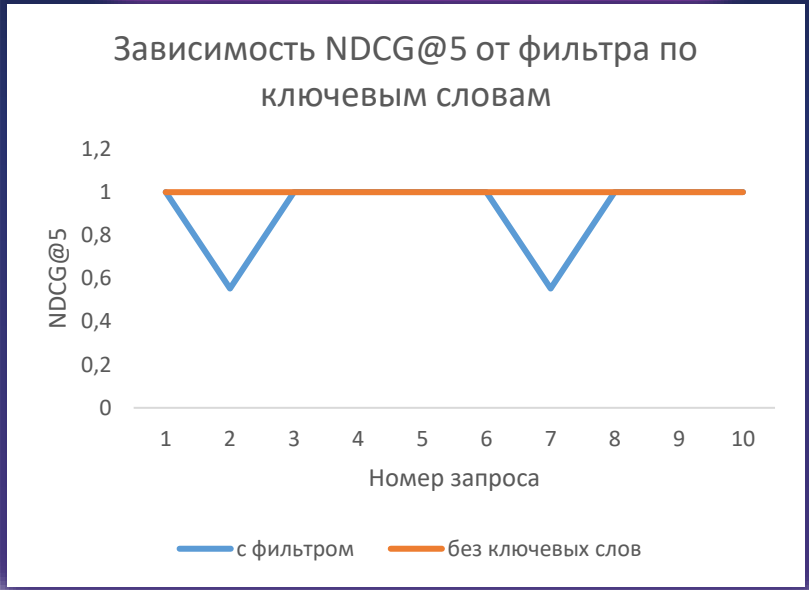
1. Проверить подключение к базе.
2. Извлечь запрос, фильтры, ключевые слова, top_k.
3. Сгенерировать эмбеддинг и сформировать фильтр.
4. Искать в Data_base_paragraphs (текст + эмбеддинг), применить фильтры, ограничить top_k.
5. Удалить дубликаты, вернуть результаты в JSON.

Архитектура решения. Бэкенд.

Анализ запросов (С тегами/ без)

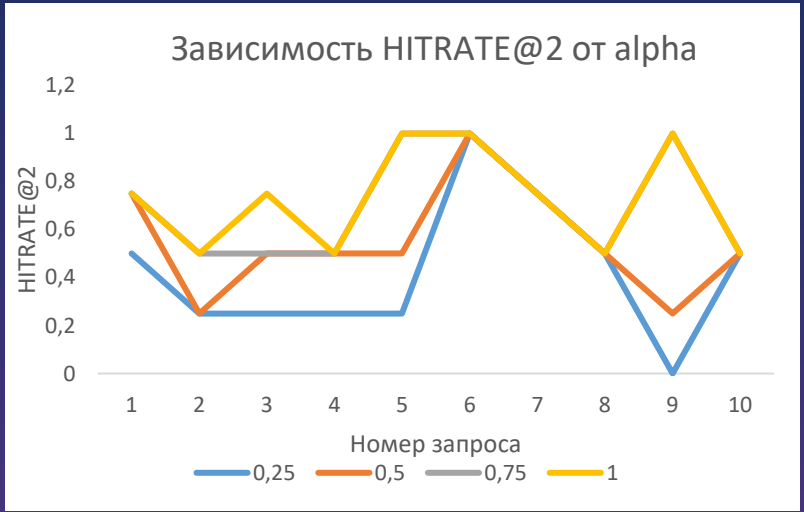


Анализ запросов (С ключевыми словами / без)

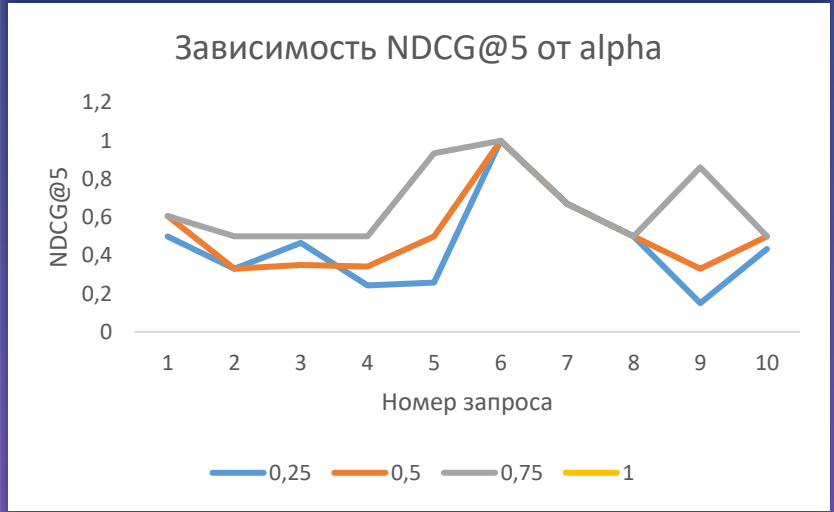


Архитектура решения. Бэкенд.

Влияние alpha на hitrate



Влияние alpha на NDCG@5



Какую языковую модель мы выбрали?

models/parameters	Embedding Stability by Model	Semantic Similarity by Model	Processing Time by Model (sec)
deepvk/USER-base	0.34	50%	0.022
BAAI/bge-m3	0.31	75%	0.027
e5-large-v2	0.35	82%	0.030
sergeyzh/LaBSE-ru-turbo	0.27	73%	0.012
sentence-transformers/all-MiniLM-L6-v2	**0.60**	**65%**	**0.010**

Команда

Кадочникова Маргарита

Капитан проекта
Data Analyst

Анализ и внедрение языковых модели, общее руководство и координация проекта

Лаврентьев Степан

Backend разработчик

Разработка клиентской части бэкенда, настройка Docker-окружения

Чикалева Юлия

Backend разработчик

Разработка FastAPI бэкенда, интеграция и настройка Weaviate, соединение фронтенда и бэкэнда

Хабаров Никита

Data Scientist
Интеграционный специалист

Оптимизация и анализ языковых моделей, оценка эффективности модели

Саркисян Арсен

Frontend разработчик
Бизнес-аналитик

Разработка UI на Streamlit, Экономический анализ и обоснование проекта, Подготовка презентации и выступления

Основные выводы

~400 000 руб.

Экономия в год с командой из 12 человек.

~80%

Сокращение времени поиска информации

5 минут

Потребуется чтобы понять как пользоваться сервисом

RAG-BOBER - Ваш интеллектуальный помощник в поиске информации

RAG

BOBER

Сервис поиска параграфов

Кадочникова Маргарита	Университет ИТМО	Прикладная математика и информатика
Лаврентьев Степан	РГУ нефти и газа (НИУ) имени И.М. Губкина	Автоматика и вычислительная техника
Чикалева Юлия	Университет ИТМО	Веб-технологии
Хабаров Никита	Университет ИТМО	Информационные системы и технологии
Саркисян Арсен	Университет ИТМО	Химия и искусственный интеллект